# VOWEL IDENTIFICATION AS INFLUENCED BY VOWEL DURATION AND FORMANT TRACK SHAPE


R.J.J.H. van Son and Louis C.W. Pols

*Institute of Phonetic Sciences, University of Amsterdam,*
*Herengracht 338, 1016 CG Amsterdam*



## ABSTRACT

*Synthetic vowels were used to investigate how listeners use vowel duration and formant track shape to determine vowel identity. The synthetic vowels had level or parabolically shaped formant tracks and variable durations. They were presented in isolation as well as in synthetic Consonant-Vowel-Consonant syllables. There was no evidence of perceptual compensatory overshoot for expected target-undershoot due to token duration or context. The only asserted effects of duration and context were in the number of long- and short-vowel responses. There was also no evidence that the listeners used the formant track shape or slopes independently to identify the synthetic vowel tokens. Tokens with curved formant tracks were mainly identified on their formant offset frequencies.*



***Keywords**: Vowel perception, perceptual-overshoot.*


## 1. INTRODUCTION

In general, theories on vowel perception assume that vowel realizations contain invariant acoustical features that allow listeners to resolve the vowel identity. One supposes that if the right transformations are performed on the acoustic signal, vowel identity will be unambiguous. Based on whether these invariant features are of a static or dynamic nature, theories on vowel perception can be divided into two "camps" [1, 5].

1) On the one side there are (elaborate) target-models that claim that the spectrum at a single cross-section in the vowel realization, i.e. the mid-point or nucleus, contains all necessary information that is used to identify it. The transition parts of the vowel realizations (i.e., the vocalic parts of CV and VC transitions) do not influence vowel recognition according to these theories, e.g. [4].

2) On the other side there are theories using dynamic-specification. These acknowledge that dynamic information from parts outside the vowel nucleus is also used to disambiguate the information from the vowel nucleus itself. Perceptual-overshoot is commonly proposed as the mechanism that compensates for expected formant-undershoot in production due to coarticulation and reduction, e.g. [2, 3].

Perceptual-overshoot is a (hypothetical) mechanism in which the listener extrapolates the course of formant transitions of the vowel on- or offset into the nucleus of the realization, overshooting the actual mid-point values realized (figure 1). In this way the listener would perceive a mid-point value closer to the canonical target of the vowel than the mid-point value actually realized acoustically. This would be a simple mechanism to achieve the aim of undoing the effects of coarticulation and reduction (i.e., formant-undershoot).

Theoretically, the shape of the formant tracks (e.g., the slope and excursion size) could be used directly to identify vowel realizations. However, the predictions obtained from using formant track shape directly or perceptual-overshoot would be indistinguishable.

A key question in this controversy is how vowel identity is affected by vowel duration and formant track shape, if it is affected at all. We could ask whether listeners do compensate for expected formant-undershoot in production and whether they use the information present in the formant transitions to perform this compensation. The alternative is that listeners would use a cross-section or some kind of formant average over (a small) part of the vowel realizations, as target-models of vowel perception predict.

The differences between models using dynamic-specification and target models seem to hinge on the effect of formant track shape on the responses of the listeners. If the vowel identity is co-specified by the formant track shape, then the targets in the responses should *overshoot* the mid-point values actually present. Furthermore, when there is real perceptual-overshoot, the amount of overshoot should depend indirectly on token *duration*, i.e. a shorter duration with steeper formant slopes should induce more overshoot. However, if formant track shape is not used to specify vowel identity, both formant track shape and duration should have *no influence* on the responses of the listeners, save some *undershoot,* due to perceptual averaging, and an exchange of long- and short-vowel responses (c.f. figure 1).

In our study we tested these predictions. We investigated how formant track shape and vowel duration influenced vowel identification, i.e. whether the responses of the listeners showed perceptual-overshoot or not. We also checked whether the presence of perceptual-overshoot depends on the presence of a non-silent *context*.

## 2. METHODS

All tokens were synthesized using an LPC-10 synthesizer with standard pre-emphasis. The (constant) synthesis parameters were: $F_0$=159, $F_3$=2490, $F_4$=3500, and $F_5$=4500 (Hz). All bandwidths were 50 Hz. Synthesis was done at 10 kHz sampling rate, 12 bit resolution, and a 4.5 kHz low-pass filter cut-off. To damp onset-transients, the synthesizer was run each time for 4 pitch-periods before actually producing

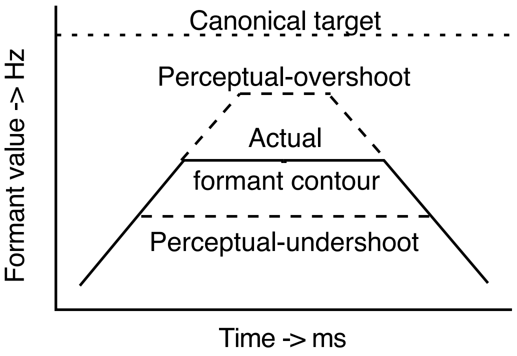


*Figure 1. Perceptual over- and undershoot. Hypothetical relation between perceived (dashed lines) versus actual formant contours (solid line). The differences are exaggerated for clarity.*

token samples. The source amplitude was constant (at 75% of maximum). To remove click sounds, the boundaries of all tokens were smoothed with a Hanning window of (2 times) 2 ms before recording. Tokens were recorded on one of the audio channels of a VCR-tape for presentation.

Nine formant "target" pairs ($F_1$, $F_2$) were defined using published values for Dutch vowels. These pairs corresponded approximately to the vowels /i u y ɩ o ɛ ɑ a œ/ and were tuned to give slightly ambiguous percepts (see table 1).

For these nine targets, smooth formant tracks were constructed for $F_1$ and $F_2$ that were either level or parabolic curves according to the following equation:

$F_n(t) = \text{Target} - \Delta F_n \cdot (4 \cdot (t/\text{Duration})^2 - 4 \cdot t/\text{Duration} + 1)$
in which:
$F_n(t)$ - the value of formant n (i.e., $F_1$ or $F_2$) at time t.
$\Delta F_n$ - the excursion size: $F_n$(mid-point) - $F_n$(on/offset).
Target - the formant target frequency.
Duration - the total token duration ($0 \leq t \leq$ Duration).

Table 1 contains the combinations of formant mid-point frequencies and excursion sizes for which formant tracks were constructed. No tracks were constructed that would cross other formant tracks or $F_0$ (indicated by '-' in table 1). All tracks were synthesized with durations of 25, 50, 100, and 150 ms. Stationary tokens with level formant tracks (i.e., $\Delta F_1 = \Delta F_2 = 0$) were also synthesized with durations of 6.3 and 12.5 ms. Of the other tokens (with either $\Delta F_1 \neq 0$ or $\Delta F_2 \neq 0$), the first and second half of the tracks (i.e., on- and off-glide only) were also synthesized with half the duration of the

*Table 1. Vowel token formant mid-point frequencies. The first three columns contain the vowel symbols and $F_1$ and $F_2$ mid-point frequencies (Hz). The other columns mark whether this mid-point frequency was used to construct tokens with the specified excursion sizes (+) or not (-).*

| V | $F_1$ | $F_2$ | $\Delta F_1$ | 0 | 225 | -225 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|
| | | | $\Delta F_2$ | 0 | 0 | 0 | 375 | -375 |
| i | 300 | 2450 | | + | - | + | + | - |
| u | 300 | 750 | | + | - | + | - | + |
| y | 300 | 1900 | | + | - | + | + | + |
| ɩ | 450 | 2200 | | + | + | + | + | - |
| o | 450 | 900 | | + | + | + | + | + |
| ɛ | 650 | 1950 | | + | + | + | + | + |
| ɑ | 700 | 1100 | | + | + | + | + | + |
| a | 750 | 1300 | | + | + | + | + | + |
| œ | 450 | 1550 | | + | + | + | + | + |

"parent" token (12.5, 25, 50, and 75 ms). The total number of tokens was 423. Some other tokens with smaller excursion sizes were used too, these will not be discussed in this paper (72 tokens, but see [6, 8]). The complete set of 495 tokens was recorded in a pseudo-random order and used in experiment 1.

A single realization each of a synthetic /n/ and /f/ sound with durations of 95 ms were obtained from a Dutch speech synthesizer. These two specific realizations were used in mixed pseudo-syllabic stimuli. Vowel tokens with durations of 50 and 100 ms and mid-point formant frequencies corresponding to /ɩ ɛ ɑ o/ (see table 1) were combined with these consonants in both /nVf/ and /fVn/ pseudo-syllables. Furthermore, the corresponding vowel tokens with only the on- or off-glide part of parabolic formant tracks (50 ms durations only) were used in CV and VC structures respectively. For comparison, corresponding *stationary* vowel tokens with 50 ms duration were also used in CV and VC pseudo-syllables. Each vowel token was presented in isolation and in these pseudo-syllables. In total, there were 220 stimuli. These were mixed with 100 similar filler tokens of the other vowels, leading to a total of 320 stimuli which were recorded twice in different pseudo-random orders and used in experiment 2.

In experiment 1, the pseudo-random sequence of 495 tokens was presented in blocks of 10 with a 3.5 s interstimulus interval to 29 Dutch subjects. Participation was voluntary. The subjects were asked to mark a vowel on an answering sheet containing an orthographic representation of all 12 Dutch monphthongal vowels (/ø œ ɛ e ɩ i y u o ɔ ɑ a/), i.e. a forced choice paradigm. For the Dutch language, presentation in orthographic form causes no ambiguities and no training was required.

In experiment 2, the two pseudo-random sequences of 320 stimuli (containing the pseudo-syllables) were both presented to a partly different group of 15 Dutch subjects. Each subject heard both sequences. Presentations of the two sequences were separated by approximately a month. In this experiment, subjects were asked to write down orthographically whatever they heard, i.e. an open response paradigm (including consonants and diphthongs).

In both experiments, the listeners first responded to a set of 10 training stimuli (not in the test set) without receiving any feedback.

## 3. RESULTS

Processing the responses of the two experiments is not easy and the way we decided to proceed may require some clarification.

The synthetic vowel-like sounds used are certainly not identified by every listener in the same way. However, the actual label used is not really important. What is important, is the difference in the identification between one set of stimuli (e.g., with level formant tracks) and another (e.g., with non-level formant tracks). Such a difference indicates how differently the stimuli are perceived.

As an example, we will compare the responses to vowel tokens with a duration of 150 ms and a $\Delta F_1$ of 225 Hz with the corresponding stationary tokens ($\Delta F_1 = \Delta F_2 = 0$). There were 6 tokens with this excursion size (see table 1). The 174 responses (29X6) to the tokens with curved formant tracks

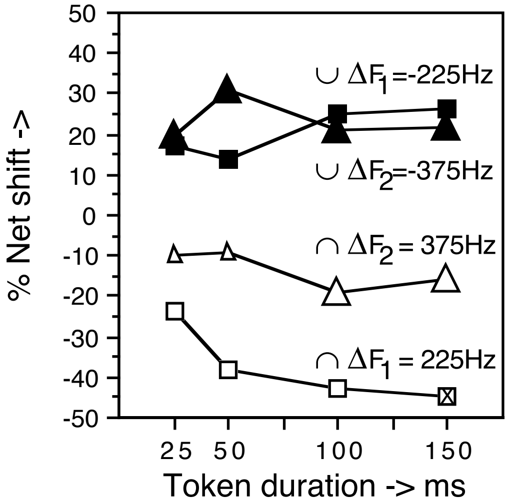


*Figure 2. Net shift in responses to tokens with curved formant tracks with respect to stationary tokens. All values were significant (p≤0.1%) except the two small open triangles. The value of the marked open square is calculated in the example in the text. Open squares: $\Delta F_1$=225Hz (∩), filled squares: $\Delta F_1$=-225Hz (∪), open triangles: $\Delta F_2$=375Hz (∩), filled triangles: $\Delta F_2$=-375Hz (∪).*

were compared to the corresponding 174 responses to the stationary tokens of equal duration, 102 of the 174 labels (59%) were different from one case to the other. Given the rather consistent fixed rank order of the Dutch vowels along the $F_1$, viz. /i y u ɩ e ø o œ ɔ ɛ ɑ a/, 90 (52%) of these responses had a lower rank number for the curved $F_1$-stimulus than for the stationary stimulus. Only 12 (7%) had a higher rank number. The net effect is thus a shift towards a lower rank number in 90-12=78 (45%) of the responses. Using a two-tailed sign-test, it is clear that this difference is statistically significant (p≤0.1%). The 45% net shift is the marked entry in figure 2, where the negative sign indicates a shift towards a lower $F_1$ rank number. The same procedure was applied to every combination of token duration, $F_1$ or $F_2$ excursion size, and shape (complete, onglide-only, offglide-only). The fixed rank order of the Dutch vowels along F2 was /u o ɔ ɑ a œ ø y ɛ e ɩ i/.

In experiment 1, token duration had a strong effect on the number of long- and short-vowel responses (figure 3). This effect was caused by the exchange of short- and long-vowel labels within the four durational oppositions of Dutch: /ɑ a:/, /ɩ e:/, /ɔ o:/, and /œ ø:/.

For the stationary tokens with a duration of 25 ms and up, 20 or more out of our 29 subjects (>67%) either used the same label for tokens with the same formant values or choose one of a long/short vowel pair. The only discrepancy between the subjects was whether some tokens represented long or short vowels. At still shorter durations, there was an increase in the number of mid-$F_1$ vowel responses (i.e., /ɩ ɔ œ/ in that order). For tokens with a duration of 6.3 ms, /ɩ ɔ/ make up almost half of the responses. Still, even for these very short tokens, at least 14 out of the 29 subjects used the same label in their responses to each individual token.

Apart from the long-short ambiguity and the rather indiscriminate increase in mid-$F_1$ vowel labels, there was no other systematic effect of duration on the responses to stationary tokens. Figure 2 contains the net-shifts in the

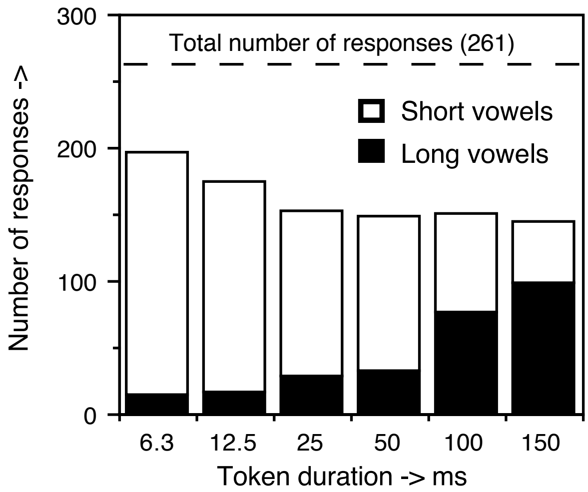


*Figure 3. The influence of token duration on the number of long- and short-vowel responses in experiment 1 (stationary tokens only). Only labels from the four long/short oppositions of Dutch were counted (i.e.,* /a: e: o: ø:/ *and* /ɑ ɩ ɔ œ/ *respectively).*

responses to tokens with curved formant tracks (complete tracks) with respect to the stationary tokens. The most salient feature of this graph is that the sign of the net-shift in the responses is always opposite to the sign of the excursion size. This means that the subjects showed "perceptual-undershoot" in their responses. That is, the responded vowel labels have mid-point formant frequencies that are shifted towards the formant on- and offset frequencies of the curved tokens with respect to the responses to the stationary tokens (which had identical mid-point frequencies).

Furthermore, the net-shift becomes smaller at shorter durations. This could be the result of "spectral blurring" at the quite short durations involved. Still, the perceptual-undershoot is present at token durations of 25 ms (4 pitch periods).

The on- and offglide-only tokens, with durations from 12.5 ms to 75 ms, showed the same pattern (not shown): A clear perceptual-undershoot that becomes smaller with smaller token durations, but remains significant down to the smallest durations (12.5 ms). In figure 4 (leftmost bars), the results combined for all durations are given. It is clear that the onglide-only tokens induced the smallest shifts and the offglide-only tokens the largest. This difference was significant for all token durations and excursion sizes (p≤0.1%, sign test). The tokens with complete parabolic formant tracks induced shifts that were in between these two.

The results of experiment 2 confirmed those of experiment 1 (see figure 4). There was a small effect of task and token mid-point values on the size of the induced undershoot. In addition, experiment 2 showed that presenting the vowel tokens in a non-integrated context only changed the number of long-vowel responses. In "closed" syllables (CVC and VC) the number of long vowel-responses was reduced, in "open" syllables (CV) it was increased. This can probably be explained by phonotactic constraints. In Dutch, open syllables cannot contain short vowels.

## 4. DISCUSSION

The fact that the offglide-only tokens always induced the largest shift in responses shows that our listeners used mostly

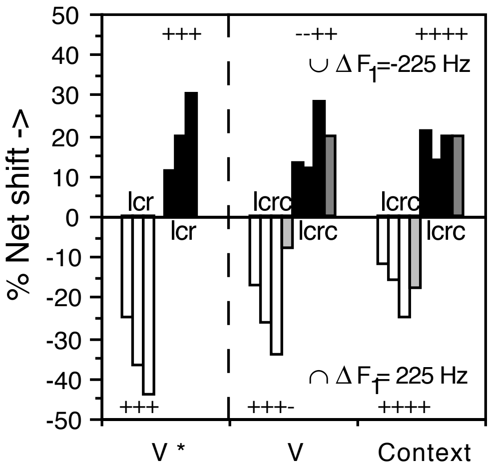


*Figure 4.a. Net shift in responses as a result of curvature of the $F_1$. 'V*' are the results of experiment 1 (all tokens pooled on duration, n>2500). 'V' and 'Context' are the results of experiment 2 with 4 mid-point pairs for tokens presented in isolation (n=120) or in context (CV, CVC, VC; C one of /n f/, n=240). Gray bars: 100 ms, white/black bars: 50 ms, l=onglide-only, c=complete, r=offglide-only tracks. +: significant (p≤0.1%, sign test), -: not significant.*

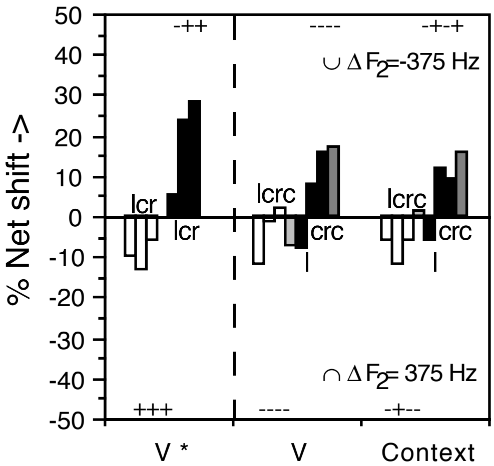


*Figure 4.b. As 4.a. but now for tokens with curved $F_2$ tracks. Dark bars: n=90 (V) or 180 (Context).*

formant values from the final part of each token to identify it. This was found for all durations and both in isolation as well as in pseudo-syllables with /n/ and /f/. The perceptual-undershoot was consistently found for all four track shapes, i.e. concave downward and upward, both for $F_1$ and $F_2$. However, the predominance of the final part of the tokens in the responses could not be found for tokens with $\Delta F_2$=375Hz (see figure 4.b).

The size of the shift in responses due to perceptual-undershoot was almost insensitive to duration. It appears that listeners used some kind of weighted average of each formant track, scaled for token duration in which most emphasis was laid on the final half of the vowel tokens. The fact that listeners were not influenced by token duration complements the fact that, in production, vowel mid-point values and excursion sizes are not changed in response to changes in vowel-duration alone [6, 7, 8].

The fact that we found that a relatively small part of each token was used to identify it would be in agreement with (compound) target models [1, 5]. However, target models assume that listeners use the vowel kernel or nucleus to identify it. In our study listeners used the offset part. The relevant literature does not supply data on how listeners detect the vowel kernel in natural speech. It is generally assumed that listeners somehow use the vowel mid-point or the part with the least spectral change. Both these strategies can be ruled out for our tokens.

To determine the role of the context in determining the perceptual mid-point, we presented vowel tokens also in pseudo-syllables (i.e., /nVf/ or /fVn/). This did not change the responses markedly. From this we can conclude that the sheer presence of (not integrated) speech surrounding a vowel will not induce compensation for coarticulation nor will it shift the "identification" point of the token towards the mid-point. As such a compensation or shift does occur in response to natural speech, other factors, like a specific and more integrated context, must be crucial in vowel perception.

Whatever the reasons for our unexpected results, they do show that current models of vowel perception are incomplete. When dynamic-specification is important in normal speech perception, factors other than the mere shape of the first and second formant track are of crucial importance. When listeners use a (compound) target, determining its position inside the vowel might be a non-trivial problem.

# 5. CONCLUSIONS

We conclude that our listeners did not use perceptual-overshoot or dynamic-specification to identify synthetic vowel tokens. Neither did they use the vowel mid-point. This implies that listeners do not automatically compensate for coarticulation or reduction at the level of the individual vowel token.

# REFERENCES

[1]: *Andruski, J.E.; Nearey, T.M.*: On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. J. Acoust. Soc. Am., Vol. 91, pp. 390-410, 1992

[2]: *Fox, R.A.*: Dynamic information in the identification and discrimination of vowels. Phonetica, Vol. 46, pp. 97-116, 1989

[3]: *Lindblom, B.; Studdert-Kennedy, M.*: On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am., Vol. 42, pp. 830-843, 1967

[4]: *Nearey, T.M.*: Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am., Vol. 85, pp. 2088-2113, 1989

[5]: *Strange, W.*: Evolving theories of vowel perception. J. Acoust. Soc. Am., Vol. 85, pp. 2081-2087, 1989

[6]: *Pols, L.C.W.; Van Son, R.J.J.H.*: Acoustics and perception of dynamic vowel segments. Speech Comm., in press 1993

[7]: *Van Son, R.J.J.H.; Pols, L.C.W.*: Formant movements of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am., Vol. 92, pp. 121-127, 1992

[8]: *Van Son, R.J.J.H.*: Spectro-temporal features of vowel segments. Thesis, University of Amsterdam, in press 1993.